Estimating Causal Effects from Nonparanormal Observational Data

Mahmoudi, Seyed Mandi; Wit, Ernst C.

Published in:
The international journal of biostatistics

DOI:
10.1515/ijb-2018-0030

Publication date:
2018

Estimating Causal Effects from Nonparanormal Observational Data

Seyed Mahdi Mahmoudi[1], Ernst C. Wit[2]

[1]Department of Statistics, Semnan University, Iran, E-mail: mahmoudi@semnan.ac.ir
[2]Bernoulli Institute, University of Groningen, Netherlands, Email: e.c.wit@rug.nl

19 June, 2018

Abstract

One of the basic aims of science is to unravel the chain of cause and effect of particular systems. Especially for large systems, this can be a daunting task. Detailed interventional and randomized data sampling approaches can be used to resolve the causality question, but for many systems, such interventions are impossible or too costly to obtain. Recently, Maathuis et al. (2010), following ideas from Spirtes et al. (2000), introduced a framework to estimate causal effects in large scale Gaussian systems. By describing the causal network as a directed acyclic graph it is a possible to estimate a class of Markov equivalent systems that describe the underlying causal interactions consistently, even for non-Gaussian systems. In these systems, causal effects stop being linear and cannot be described any more by a single coefficient. In this paper, we derive the general functional form of a causal effect in a large subclass of non-Gaussian distributions, called the non-paranormal. We also derive a convenient approximation, which can be used effectively in estimation. We show that the estimate is consistent under certain conditions and we apply the method to an observational gene expression dataset of the Arabidopsis thaliana circadian clock system.

1 Introduction

Inferring cause-and-effect relationships between variables is of primary importance in many fields of science. The classical approach for determining such relationships uses randomized experiments where a single or few variables are perturbed.

Such intervention experiments, however, can be very expensive, unethical (e.g. one cannot force a randomly selected person to smoke many cigarettes a day) or even infeasible. Hence, it is desirable to infer causal effects from so-called observational data obtained by observing a system without subjecting it to interventions. Although some important concepts and ideas have been worked out (Spirtes, Meek, and Richardson, 1995, Richardson, 1996, Mooij, Janzing, Heskes, and Schölkopf, 2011), estimating causal effects for non-Gaussian observational systems is still in its infancy.

Pearl (1995, 2009) described a do-calculus of causal effects, if the underlying causal diagram is known. In practice, though, the influence diagram is often not known and one would like to infer causal effects from observational data together with the influence diagram. Spirtes, Glymour, and Scheines (2000) introduced methods to estimate causal graphs from observational data. Verma and Pearl (1990) found that typically groups of causal graphs give rise to the same distribution of the data, which implies that the generating causal DAG is typically unidentifiable from the data. This Markov equivalence class of causal DAGs has been called completed partially directed acyclic graph (CPDAG). A CPDAG can be estimated in various ways, including the PC-algorithm (Spirtes et al., 2000), search and score methods (Chickering, 2002, 2003, Verma and Pearl, 1990) and Bayesian methods (Heckerman and Geiger, 1995, Spiegelhalter, Dawid, Lauritzen, and Cowell, 1993).

The PC-algorithm uses conditional independence tests to infer a CPDAG from data (Spirtes et al., 2000). Sample partial correlations derived from independent multivariate normal observations have favourable distributional properties (Anderson, 2003, Chapter 4), which form the basis for the work of Kalisch and Bühlmann (2007), who treat the PC-algorithm in the Gaussian context with conditional independence tests based on sample partial correlations. They prove the high-dimensional consistency of the PC-algorithm, when the observations form a sample of independent normal random vectors that are faithful to a suitably sparse DAG. Maathuis, Kalisch, and Bühlmann (2009) propose a method that combines the estimation of the causal structure and the interventional distribution in the Gaussian case. Due to the Gaussian structure, they find that the causal effects can be described by a set of regression coefficients. Harris and Drton (2013) show that the PC-algorithm has high-dimensional consistency properties for a broader class of distributions, when standard Pearson-type empirical correlations are replaced by rank-based measures of correlations in tests of conditional independence, such as Spearman's rank correlation and Kendall's tau. A special class of non-Gaussian distributions is constituted by the Gaussian copula, or, in the terminology of Liu, Han, Yuan, Lafferty, and Wasserman (2012), the so-called "nonparanormal distributions." Teramoto, Saito, and Funahashi (2014) uses this class to estimate the underlying causal DAG for the design of efficient intervention experiments. Nandy, Maathuis, and Richard-

son (2017) applied *Intervention-calculus when the DAG is Absent* (IDA) to nonparanormal distributions, and summarize the causal effects among the underlying Gaussian random variables. What is missing up until now is a way to describe and estimate the causal effects in such nonparanormal scenarios. The main difficulty is that causal effects in non-Gaussian scenarios stop being constant and become functions of the intervention variables.

In the remainder of the paper, we will consider an observational setting of a nonparanormal system. We assume the causal CPDAG has been estimated by, e.g., the Rank PC (RPC) algorithm (Harris and Drton, 2013), i.e., the PC-algorithm in the nonparanormal context. Based on the estimated CPDAG, it is our aim to derive an expression for the causal effect in this system and to find a consistent way to estimate them. In section 2, we introduce the causal graph terminology, a short description of the intervention calculus and the definition of a causal effect. In section 3, we derive the structure of a causal effect of a nonparanormal causal effect and in section 4, we define a convenient estimator. In section 5, we evaluate the performance of our method in a simulation study. Finally, in section 6, we illustrate the method in a real data example.

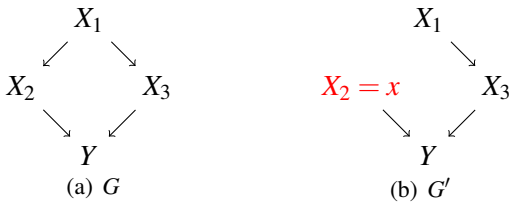
2 Causal effects in causal graphs

In this section we describe the background needed in order to define the notion of a causal effect. We begin by defining causal models through directed graphical models.

A *graph* is a pair $G = (V, E)$, where V is a finite set of *vertices* $V = \{1, 2, \ldots, p\}$, also called *nodes*, of G and E is a subset of $(V \times V)$ of ordered pairs of vertices, called the *edges* or *links* of G. We consider p random variables $X_1, \ldots, X_p$, associated to the vertices. If edge $(X_i, X_j) \in E$ but $(X_j, X_i) \notin E$, we call the edge *directed* or an *arrow*, denoted by $X_i \to X_j$. In that case, we also say that X_i is a *parent* of X_j, and that X_j is a *child* of X_i. The set of parents of a vertex X_j is denoted by $\text{pa}(j)$. We use the short-hand notation $X_i - X_j$ that is *undirected edge* to denote $(X_i, X_j) \in E$ and $(X_j, X_i) \in E$. A graph containing only directed edges ($\to$) is *directed*, one containing only undirected edges ($-$) is *undirected*. A directed graph is called a *directed acyclic graph* (DAG) if it does not contain directed cycles. A DAG of p random variables $X_1, \ldots, X_p$ can be interpreted as a Markov independence graph, describing a multivariate distribution. Various DAGs can lead to the same distribution. A common tool for describing such Markov equivalence class of DAGs are completed partially directed acyclic graphs (CPDAGs).

Pearl (2009) defined causality through intervention, whereby variables are externally manipulated to take certain values. This intervention changes the under-

lying distribution P and can be expressed by adapting the DAG. The new distribution is called the *intervention distribution* and we say that the variables, whose structural equations we have replaced have been "intervened on." The intervention distribution of Y when doing an intervention and setting the variable X_i to a value x_i' is denoted by $P(Y|\,\mathrm{do}(X_i = x_i'))$. The intervention on variable X_i is characterized by a *truncated factorization*, in which an intervention DAG G', arising from the non-intervention DAG G can be defined by deleting all edges which point into the node X_i. Consider the example graph below, a DAG G and its corresponding intervention graphs (G') are shown.



(a) G (b) G'

Figure 1: (a) A DAG G and (b) its corresponding intervention graph G'. The intervention is $\mathrm{do}(X_2 = x)$, described by the red label in the graph. The parental set of $i = 2$ is $\mathrm{pa}(2) = \{1\}$ which appears in (3) for computing the causal effect β_2 of X_2 on Y.

The *total causal effect* of X_i on Y at x_i is the relative amount Y is expected to change as a result from a small *interventional* change of X_i at x_i,

$$\mathrm{CE}(Y|X_i = x_i) = \frac{\partial}{\partial x} E[Y|\,\mathrm{do}(X_i = x)]|_{x=x_i}, \tag{1}$$

where we have that if $Y \notin X_{\mathrm{pa}(i)}$,

$$E(Y|\mathrm{do}(X_i = x)) = \int E(Y|X_i = x, X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)})\, P(x_{\mathrm{pa}(i)})\, d(x_{\mathrm{pa}(i)}). \tag{2}$$

If $(X_1, ..., X_{p-1}, Y)$ has a multivariate Gaussian distribution, then using the fact that the conditional expectation $E(Y|X_i = x, X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)})$ is linear in x_i and $x_{\mathrm{pa}(i)}$ if $Y \notin X_{\mathrm{pa}(i)}$, it is straightforward to see that

$$E(Y|\mathrm{do}\,(X_i = x_i)) = \beta_i x_i + \int \beta_{\mathrm{pa}(i)}^T x_{\mathrm{pa}(i)} P(x_{\mathrm{pa}(i)})\, d(x_{\mathrm{pa}(i)}),$$

for some coefficients β_i and $\beta_{\mathrm{pa}(i)}$. Therefore, the causal effect is given by

$$\mathrm{CE}(Y|X_i = x_i) = \frac{\partial}{\partial x} E[Y|\,\mathrm{do}(X_i = x)]|_{x=x_i} = \beta_i. \tag{3}$$

From (3), it follows that the total causal effect of X_i on Y with $Y \notin X_{\mathrm{pa}(i)}$ is given by the regression coefficient of X_i in the regression of Y on X_i and $\mathrm{pa}(i)$. Note that if $Y \in X_{\mathrm{pa}(i)}$, the total causal effect from X_i to Y is, obviously, zero. Our aim is to generalize this to a wider class of distributions.

3 Causal effect for nonparanormal graphical models

Spirtes et al. (2000) introduced the PC-algorithm to estimate causal graph from observational data. Kalisch and Bühlmann (2007) proved consistency of the PC-algorithm in a Gaussian setting for estimating the causal skeleton and, subsequently, the Markov equivalence class of high-dimensional causal graphs. The algorithm is based on a clever hierarchical scheme for testing conditional independence among pairs of variables X_j, X_k (for all $j \neq k$) in the DAG. In Gaussian models, tests of conditional independence can be based on Pearson correlations and high-dimensional consistency results have been obtained for the PC-algorithm in this setting. Harris and Drton (2013) proved high-dimensional consistency properties for a broader class of nonparanormal models when using rank-based measures of correlation. They showed that the *Rank PC-algorithm* (RPC) works as well as the Pearson PC-algorithm for normal data and considerably better for non-Gaussian data. If one assumes to know all conditional independencies exactly – the oracle setting– then the RPC-algorithm yields the "true" CPDAG, i.e., the Markov equivalence class of DAGs that contains the true causal DAG.

Building on this work in the Gaussian setting, Maathuis et al. (2009) derived an expression for and an estimator of the total causal effect of a covariate X_i on a response Y in a Gaussian causal graph. After obtaining the CPDAG Markov equivalence class of, say, m causal DAGs, they apply for each DAG G_j in this class the intervention calculus to obtain the total causal effect β_{ij} of X_i on Y. Then they define multi-sets $\Theta_i = \{\beta_{ij}\}_{j\in\{1,\ldots,m\}}$ containing the estimated possible causal effects of X_i on Y.

In this section, we derive the analogous multi-set of causal effects the *nonparanormal* setting. In practice, the conditional independences have to be inferred from the data as well and we show how using our main result in combination with the RPC-algorithm we are able to define an convenient estimator for the causal effect for such data, which stops being linear and needs to be estimated functionally.

3.1 General expression of nonparanormal causal effect

Liu et al. (2012) define the nonparanormal distribution. Let $f = (f_i)_{i\in\mathbf{V}}$ be a set of monotone, univariate functions and let $\Sigma \in \mathbb{R}^{\mathbf{V}\times\mathbf{V}}$ be a positive definite covariance

matrix. We say a p-dimensional random variable $X = (X_1, ..., X_p)^{\mathbf{T}}$ has a nonparanormal distribution,

$$X \sim \text{NPN}(\mu, \Sigma, f),$$

if $f^{-1}(X) = (f_1^{-1}(X_1), \ldots, f_p^{-1}(X_p)) \sim N(\mu, \Sigma)$. If $X \sim \text{NPN}(\mu, \Sigma, f)$, then the univariate marginal distribution for a coordinate, say X_i, can have any distribution F_i, as we can take $f_i = F_i^{-1} \circ \Phi_{\mu_i, \sigma_i^2}$, where Φ_{μ_i, σ_i^2} is the normal distribution function with mean μ_i and variance $\sigma_i^2 = \Sigma_{ii}$. Note that, in general, f_i need not be continuous. However, in this paper, we deal with monotone and differentiable f. Liu et al. (2012) show that in that case the nonparanormal distribution $\text{NPN}(\mu, \Sigma, f)$ is a Gaussian copula.

In the remainder of the paper, we assume that $(X_1, \ldots, X_{p-1}, Y) \sim \text{NPN}(0, \Sigma, f)$, where Σ is a correlation matrix. We will refer to the latent standard normally distributed variables as $Z_i = f_i^{-1}(X_i) = \Phi^{-1} \circ F_i(X_i)$ and $Z = f_y^{-1}(Y) = \Phi^{-1} \circ F_y(Y)$. We are interested in the total causal effect of X_i on Y for $i \in (1, \ldots, p-1)$. We know from section 2 that for Gaussian data it is very simple to compute the total causal effect, since Gaussianity implies that $E(Y|X_i = x_i; X_{-i} = x_{-i})$ is linear in x_i. Unfortunately, this is no longer true for non-Gaussian random variables. In Theorem 1 we derive the explicit functional form for the total causal effect in the entire class of nonparanormal distributions.

Theorem 1. *Let $(X_1, \ldots, X_{p-1}, Y) \sim NPN(0, \Sigma, f)$ and f_i $(i = 1, \ldots, p-1)$ is differentiable and f_y is infinitely differentiable, then the total causal effect of X_i on Y in causal graph G is given by*

$$\begin{aligned} CE(Y|X_i = x_i) = \sum_{k=1}^{\infty} \sum_{r=0}^{\lfloor \frac{k-1}{2} \rfloor} \sum_{s=1}^{k-2r} f_y^{(k)}(z_0) \frac{1}{k!} \binom{k-2r}{s} \binom{k}{2r} s\beta_i \\ \times (-z_0 + \beta_i z_i)^{s-1} E[(\beta_{pa(i)}^T Z_{pa(i)})^{k-2r-s}] \\ \times (2r+1) \times \ldots \times 3 \times 1 \times [(1-\rho^2)]^r (f_i^{-1})'(x_i), \end{aligned} \tag{4}$$

for every $z_0 \in \mathbb{R}$, where $f_y^{(k)}$ is the kth derivative of f_y, $z_i = f_i^{-1}(x_i)$, $Z_{pa(i)} = f_{pa(i)}^{-1}(X_{pa(i)})$, $(\beta_i, \beta_{pa(i)}) = \Sigma_{p,(i,pa(i))} \Sigma^{-1}_{(i,pa(i)),(i,pa(i))}$ and $\rho = (\beta_i, \beta_{pa(i)}) \Sigma_{(i,pa(i)),p}$.

The proof of the theorem is given in the appendix. We have obtained the general expression (4) for a nonparanormal causal effect. The value of this theorem is that it gives us insight in how higher order moments of the effect Y, captured in the higher order derivatives of f_y, affect the causal effect, whereas higher order moments of the cause X_i do not. In practice, this formula is not very helpful as it contains information about the system that we typically do not possess, such as

the correlation structure of the latent normal variable. However, this formula can inspire practical estimation procedures of the causal effects in nonparanormal systems. Whereas this is in principle possible, we restrict our attention in this paper to a lower order Taylor approximations. A first and second order Taylor expansion are given by

$$\begin{aligned} CE_{1,z_0}(Y|X_i = x_i) &= f_y'(z_0)\beta_i(f_i^{-1})'(x_i), \\ CE_{2,z_0}(Y|X_i = x_i) &= \left[f_y'(z_0)\beta_i + f_y''(z_0)\beta_i(\beta_i z_i - z_0)\right] \times (f_i^{-1})'(x_i). \end{aligned}$$

If median and mode of Y coincide, then it is easy to show that the second order Taylor expansion collapses to the first order by taking $z_0 = 0$, i.e., $CE_{2,0}(Y|X_i = x_i) = (F^{-1})'(0.5)\beta_i(f_i^{-1})'(x_i)$. Higher order expansions become quickly more intricate. Moreover, especially when it comes to estimation in section 4, the estimates involved in lower order expansions tend to be intrinsically more stable.

3.2 Special case

We consider the special case of the above theorem for the situation that only Y is normally distributed, and the X_is are still nonparanormal.

Corollary 1. *Let $(X_1, \ldots, X_{p-1}) \sim NPN(0, \Sigma, f)$ and f_i $(i = 1, \ldots, p-1)$ is differentiable and $Y \sim N(\mu, \sigma^2)$, then the total causal effect of X_i on Y in causal graph G is given by*

$$CE(Y|X_i = x_i) = \sigma\beta_i(f_i^{-1})'(x_i), \tag{5}$$

where β_i is defined as in Theorem 1.

The result simply follows from $f_y(Z) = \mu + \sigma Z$ for Z standard normal. This special case both inspires an estimator for the causal effect and gives some hope for obtaining some consistency results.

4 NCE: nonparanormal causal effect estimator

In this section, we propose a simple estimator for the causal effect that is able to capture non-linear effects for a wide-ranging collection of distributions. Furthermore, we show that under some conditions, this estimator is consistent.

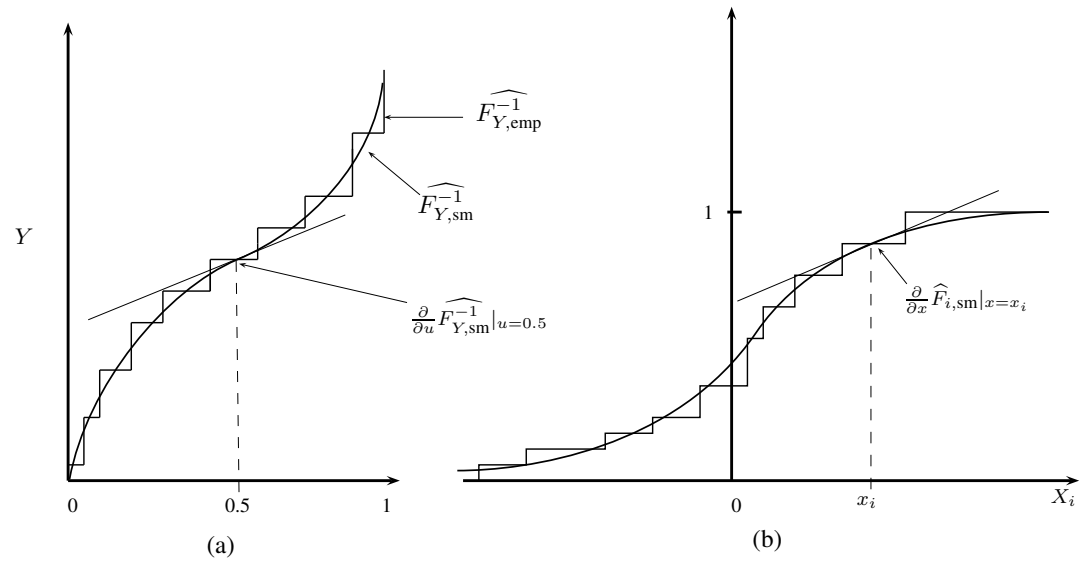


Figure 2: (a) the derivative of monotone increasing spline $\widehat{F^{-1}_{Y,\text{sm}}}$ for estimate $\frac{\partial}{\partial x}F_Y^{-1}$. (b) the derivative of the monotone increasing estimating spline $\widehat{F}_{i,\text{sm}}$ for estimate $\frac{\partial}{\partial x}F_i$.

4.1 First order estimator

In section 3.2, we derived a one-term expression that is used as inspiration for a first-order Taylor estimator of the general causal effect of $X_i = x$ on Y, i.e.,

$$\widehat{\text{NCE}}_{z_0}(x) = \hat{f}_y'(z_0)\ \hat{\beta}_i\ (\widehat{f_i^{-1}})'(x), \tag{6}$$

for some z_0, $x \in \mathbb{R}$ and where $\hat{\beta}_i$ is the linear regression coefficient of $\widehat{f_y^{-1}}(Y)$ on $\widehat{f_i^{-1}}(X_i)$, while controlling for the parents $\widehat{f^{-1}_{\text{pa}(i)}}(X_{\text{pa}(i)})$ of i, given some estimators for the various functions f^{-1} and f_y. We will show in the next section that in order to obtain consistency, we trim the data for each variable below its α/p and above $1-\alpha/p$ quantiles, where p is the number of random variables (X,Y). When an observation has been trimmed for one variable, it is removed in its entirety for all variables. This means that in the worst case scenario, $1-2\alpha$ of the observations remain. In practice, we will often use $\alpha = 0.05$.

We can simplify expression (6) by considering the case that $z_0 = 0$. Note that

it is straightforward to obtain

$$f_y^{'}(0) = \frac{\partial}{\partial u} F_Y^{-1}(u)|_{u=0.5}\ \phi(0)$$
$$(f_i^{-1})'(x) = \left[\phi(f_i^{-1}(x))\right]^{-1}\ \frac{\partial}{\partial x} F_i(x),$$

where ϕ is the density function of a standard normal distribution. Considering Figure 2, F_Y^{-1} will be estimated via a monotone increasing smoother $\widehat{F_{Y,\mathrm{sm}}^{-1}}$, which gives us direct access to its derivative. Similarly, $\frac{\partial}{\partial x} F_i$ will be estimated by taking the derivative of the monotone increasing estimating smoother $\widehat{F}_{i,\mathrm{sm}}$. Finally, $f_i^{-1}(x)$ will be estimated as $\hat{z} = \Phi^{-1}(\hat{F}_{i,\mathrm{sm}}(x))$. Putting this together, we obtain a simplified and explicit estimator of a non-paranormal causal effect,

$$\widehat{\mathrm{NCE}}_0(x) = \hat{\beta}_i\ \frac{\phi(0)}{\phi(\hat{z})}\ \frac{\widehat{\partial F_{Y,\mathrm{sm}}^{-1}}}{\partial u}(0.5)\ \frac{\partial \widehat{F}_{i,\mathrm{sm}}}{\partial x}(x). \tag{7}$$

In the following section, we will show that under certain conditions the above estimator is consistent. In particular, we show that estimating F_Y^{-1} and F_i with a particular kind of kernel smoother works. Nevertheless, in practice any slightly stiff smoother will result in almost the same estimates. A natural cubic spline is easy to implement and can be easily differentiated, which is needed for the causal effect estimator $\widehat{\mathrm{NCE}}_0(x)$.

4.2 Consistency

In this section we will be concerned with the asymptotic behaviour of our estimator in (7) under the assumption of normality of Y, but no such assumption on the X_i. We first show that the random, but not necessarily independent, sampling scheme of $(X_1,\ldots,X_{p-1}) \sim NPN(0,\Sigma,f)$ and $Y \sim N(\mu,\sigma^2)$ combined with our lower and upper α/p trimming scheme will eventually fill up the p-dimensional cube $[L_\alpha, U_\alpha]$, where $L_\alpha = (L_\alpha^1,\ldots,L_\alpha^{p-1},L_\alpha^y)$ and $U_\alpha = (U_\alpha^1,\ldots,U_\alpha^{p-1},U_\alpha^y)$ are the α- and $(1-\alpha)$-quantiles , respectively, for each of the variables $(X_1,\ldots,X_{p-1},Y)$. From the original sample size n approximately $(1-2\alpha)n$ will fall in this cube. Then we show that the kernel estimators of the functions used in the NCE estimators and their derivatives converge fast to their true values in probability. Together with the fact that products of consistent estimators are consistent, this proves the consistency of the estimator $\widehat{\mathrm{NCE}}_0(x)$.

Proposition 1. *Consider any absolutely continuous random variable X with lower α quantile L_α and upper α quantile U_α. For the $N \asymp (1-2\alpha)n$ ordered observations of X in the finite interval $[L_\alpha, U_\alpha]$, the following property holds*

$$\max_{2\le i\le N} |X_{(i)} - X_{(i-1)}| = O_P(1/N).$$

The symbol $\asymp$ denotes that two sequences of real numbers are asymptotically of the same order. The proof of this Proposition is a simple exercise and will not be given here.

Our goal is first to estimate the function F_i and its derivative $\frac{\partial}{\partial x}F_i$. Similarly, we aim to estimate F_i^{-1} and its derivative. In order to derive asymptotic properties, we will be using kernel estimators for $\hat{F}_{i,sm}$ and $\widehat{F^{-1}_{i,sm}}(x)$, respectively,

$$\hat{F}_{i,n}(x) = \sum_{j=2}^{N} (x_{i(j)} - x_{i(j-1)}) \frac{1}{b_n} K\left(\frac{x - x_{i(j)}}{b_n}\right)\left(\alpha + \frac{j-1}{n}\right), \tag{8}$$

$$\widehat{F^{-1}_{i,n}}(u) = \sum_{j=1}^{N} \frac{1-2\alpha}{N} \frac{1}{b_n} K\left(\frac{u - (\alpha + \frac{j(1-2\alpha)}{N})}{b_n}\right) x_{i(j)}, \tag{9}$$

for $x \in [L^i_\alpha, U^i_\alpha]$ and $u \in [\alpha, 1-\alpha]$, where K is a kernel function, (Priestley and Chao, 1972), $b_n > 0$ denotes the bandwidth that we take to depend on the sample size n in such a way that $b_n \to 0$ as $n \to \infty$ and $x_{i(1)}, x_{i(2)}, \ldots, x_{i(N)}$ denote the order statistics of that part that for the i variable that falls within $[L^i_\alpha, U^i_\alpha]$. Note that we have selected the same bandwidth for the quantile function and the CDF, even though they could live on completely different domains. In practice, it may be sensible to scale the bandwidth by some constant depending on the domain. However, for proving consistency we do not need it. We define an estimator of $\frac{\partial}{\partial x}F_i$ by taking the derivative of the kernel smoother $\widehat{\frac{\partial}{\partial x}F_{i,n}} = \frac{\partial}{\partial x}\hat{F}_{i,n} = \hat{F}'_{i,n}$.

Proposition 2. *If the kernel K is symmetric and twice continuously differentiable with support in $[-1,1]$ and if it satisfies the integrability conditions (a) $\int_{-1}^{1} K(u)\,\mathrm{d}u = 1$ and (b) $\int_{-1}^{1} u^\ell K(u)\,\mathrm{d}u = 0$ for $\ell = 1, \ldots, \gamma - 1$, then for a fixed number δ, such that $\alpha < \delta < 1/2$:*

(i) if F and F^{-1} are $\gamma \ge 1$ times continuously differentiable and $b_n \to 0$ as $n \to \infty$, then

$$\sup_{x \in [L^i_\alpha, U^i_\alpha]} |\hat{F}_{i,n}(x) - F_i(x)| = \mathrm{O}_P\left(b_n^\gamma + \frac{1}{nb_n^2} + \sqrt{\frac{\log n}{nb_n}}\right).$$

$$\sup_{u \in [\delta, 1-\delta]} |\widehat{F^{-1}_{i,n}}(u) - F_i^{-1}(u)| = \mathrm{O}_P\left(b_n^\gamma + \frac{1}{nb_n^2} + \sqrt{\frac{\log n}{nb_n}}\right).$$

(ii) If F and F^{-1} are $\gamma \geq 2$ times continuously differentiable and $b_n \to 0$ as $n \to \infty$, then

$$\begin{aligned}
\sup_{x \in [L_\alpha^i, U_\alpha^i]} |\hat{F}'_{i,n}(x) - F_i'(x)| &= \mathrm{O}_P\left(b_n^{\gamma-1} + \frac{1}{nb_n^3} + \sqrt{\frac{\log n}{nb_n^3}} \right). \\
\sup_{u \in [\delta, 1-\delta]} |\widehat{F_{i,n}^{-1}}'(u) - F_i^{-1\prime}(u)| &= \mathrm{O}_P\left(b_n^{\gamma-1} + \frac{1}{nb_n^3} + \sqrt{\frac{\log n}{nb_n^3}} \right).
\end{aligned}$$

In particular, $\hat{F}_{i,n}(x)$ and $\hat{F}'_{i,n}(x)$ are consistent on $[L_\alpha^i, U_\alpha^i]$ and $\widehat{F_{i,n}^{-1}}(x)$ and $\widehat{F_{i,n}^{-1}}'(x)$ are consistent on $[\delta, 1-\delta]$, if $nb_n^3/\log n \to \infty$ holds additionally.

The proof is given in Gugushvili and Klaassen (2012, Proposition 3.1). The estimator $\widehat{\mathrm{NCE}}_0(x)$ in (7) contains four terms. Based on Proposition 2 the two terms $\hat{F}'_{i,n}(x)$ and $\widehat{F_{i,n}^{-1}}'(x)$ are consistent. As any continuous function of a consistent estimator is consistent (Lehmann, 2004), also $\hat{z} = \Phi^{-1}(\hat{F}_{i,\mathrm{n}}(x))$ is consistent. In order to proof consistency of $\widehat{\mathrm{NCE}}_0(x)$ we still need to show that $\hat{\beta}_i$ is consistent, where $\hat{\beta}_i$ is the linear regression coefficient of $\widehat{f_y^{-1}}(Y)$ on $\widehat{f_i^{-1}}(X_i)$, while controlling for the parents $\widehat{f_{\mathrm{pa}(i)}^{-1}}(X_{\mathrm{pa}(i)})$ of i. The following proposition shows the consistency of $\hat{\beta}_i$.

Proposition 3. *Let $\hat{\beta}_i$ be the linear regression coefficient of $\widehat{f_y^{-1}}(Y)$ on $\widehat{f_i^{-1}}(X_i)$, while controlling for the parents $\widehat{f_{pa(i)}^{-1}}(X_{pa(i)})$ of i, then*

$$\hat{\beta}_i^n \xrightarrow{P} \beta_i, \tag{10}$$

where β_i is the true regression coefficient as defined in Theorem 1.

The formal proof is given in the appendix and is again based on the fact that $\widehat{f_y^{-1}}(Y)$, $\widehat{f_i^{-1}}(X_i)$ and $\widehat{f_{\mathrm{pa}(i)}^{-1}}(X_{\mathrm{pa}(i)})$ are consistent estimators and $\hat{\beta}_i^n$ is a continuous function of these consistent estimators and, therefore, consistent. Putting the previous results together we can now show that our estimator $\widehat{\mathrm{NCE}}_0(x)$ in (7) is consistent. The proof is given in the appendix.

Proposition 4. *Consider the estimator of $NCE_0(x)$ in (7), for which we consider the component estimators (8), (9) and (10). For the kernel estimators, we assume that the conditions of Proposition 2 are satisfied and, furthermore, the bandwidth $b_n \to 0$, but not too fast so that $nb_n^3/\log n \to \infty$. Then we have*

$$\widehat{NCE}_{0,n} \xrightarrow{\mathrm{P}} NCE_0.$$

5 Simulation studies

In this section, we test our NCE estimation method for two different types of distributions, to wit, Gaussian and standard Cauchy. For Gaussian data, the method should find constant causal effects and can be compared directly with the IDA method (Maathuis et al., 2009). We consider two scenarios: (i) in which the underlying causal graph is known and (ii) where it is unknown and needs to be estimated via the RPC-algorithm. Secondly, we show how our NCE method captures the non-linear nature of causal effects for a bivariate exponential system. Finally, we compare the NCE causal graph reconstruction method, based on the RPC algorithm, with the nonparanormal (NPN) method by Teramoto et al. (2014) in a system with Cauchy distributed data.

5.1 Gaussian data

Following Kalisch and Bühlmann (2007), we simulate random DAGs and sample from probability distributions faithful to them. For convenience, we fix an increasing ordering of the variables $\{X_1, \ldots, X_p\}$, meaning that for a vector of independent Gaussian variables $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$

$$X = BX + \varepsilon, \tag{11}$$

where the lower diagonal coefficient matrix B has entries β_{ij} that are zero for $i < j$ and $\beta_{ji} \neq 0$ if the corresponding DAG has a directed edge from node i to node j for some $i > j$. The entries β_{ij} are by definition the causal effects of X_j on X_i. We create a DAG G with an expected vertex degree of three by drawing edges (i, j) for $i > j$ independently with probability $3/p$. The nonzero entries of B are drawn from independent standard normal distributions. Then, with probability one, the vector X solving (11) is Markov and faithful with respect to G. We consider two different size graphs: a small graph with $p = 10$ vertices and a larger graph with $p = 50$ vertices. For each $n \in \{100, 1000\}$ and each of the two types of graphs, we repeat the simulation 100 times.

5.1.1 Causal DAG known

If we assume that the causal DAG is known, then for estimating the causal effects we apply both our NCE algorithm, described in section 4 and the IDA algorithm by Maathuis et al. (2009), which estimates β_{ij} via least squares linear regression,

$$x_i = \beta_{i0} + \beta_{ij} x_j + \beta_{i,\mathrm{pa}(j)} x_{\mathrm{pa}(j)}.$$

Given that the IDA algorithm is made for these Gaussian data, the method should outperform the NCE method, which is agnostic about the underlying distributional assumptions. We apply the methods to the four data scenarios and the results are presented in the last column of Table 1 as mean absolute deviations. The mean absolute deviation is a robust version of the mean squared error and smaller values refer to better estimates of the causal effects. Table 1 shows that when the number of observations are increasing, the mean absolute deviation for causal effect estimates for both IDA and NCE methods is decreasing. Furthermore, the NCE method, as expected, is somewhat more variable. This variation is mostly the result from the poorer estimates of the distributional shape in the tails of the distribution.

5.1.2 Causal DAG unknown

If the underlying causal DAG is considered unknown, then the CPDAG and associated DAGs need to be estimated. For each simulation, we run both the standard PC-algorithm and the robust RPC-algorithm on a grid of significance levels α ranging from 10^{-10} to 0.5. For each estimated DAG, we compute the causal effects of each node according to the NCE method and the compare the results with the IDA method.

Figures 3 show the causal effects between the chosen nodes for small graph on ten vertices $p = 10$ with $n = 100$. In these figures the red line show the real causal effect between two chosen nodes. The blue line shows the average estimated causal effect from the IDA method. The black line show the average functional causal effect estimate (7) proposed by our NCE method across the DAGs consistent with the inferred CPDAG. The dashed lines express the average standard deviation of our functional causal effect estimate. A clear message emerges from plots: whereas the IDA method is exactly matched for this simulation scenario, our nonparanormal causal effects estimates are quite stable. Moreover, the confidence intervals calculated by our method typically contain the true effect.

In Table 1 provide numerical comparisons of both methods on data sets with different transformations, where we repeat the experiments 100 times and report the mean absolute deviation for causal effect estimates on each pair nodes in both IDA and NCE methods. Even though the simulation method is precisely suited for the IDA method, our NCE method is highly competitive.

5.2 Exponential data

Only in a few special non-Gaussian distributional examples can we calculate the causal effects exactly. This makes large scale simulation studies difficult. There-

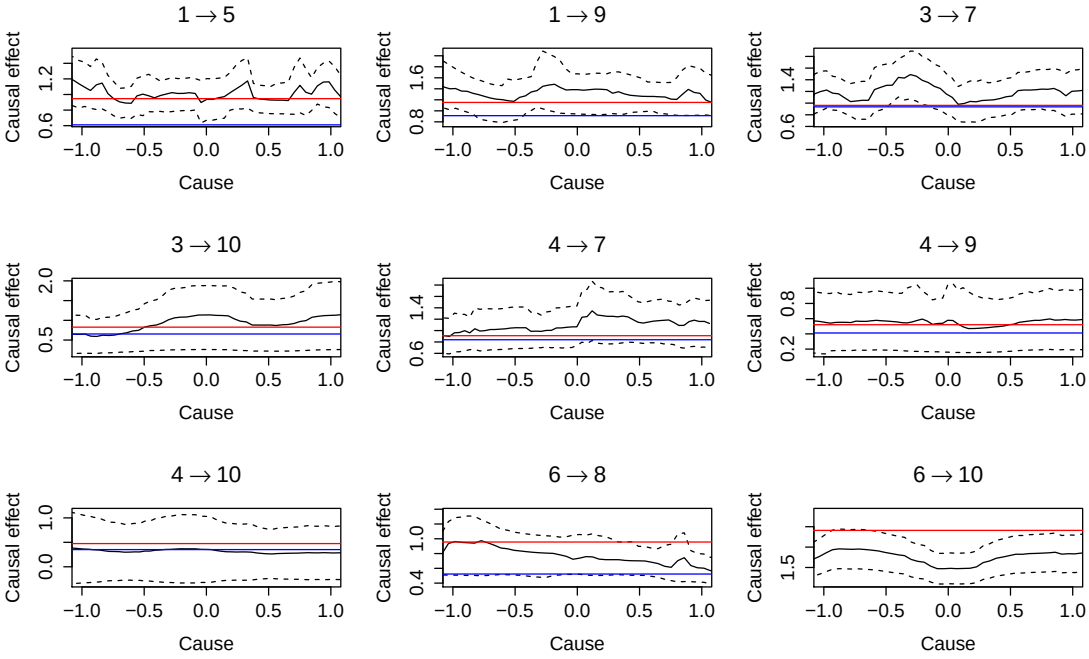


Figure 3: Simulation study for Gaussian data from a causal graph ($p = 10$ vertices with $n = 100$ observations). The red lines are the true (constant) causal effects. The blue lines are the causal effect estimates from the IDA methods and black lines show the functional causal effect estimates from our NCE method. The dashed lines show the confidence intervals for functional NCE causal effect estimates.

fore, in this section we consider the causal effects in a bivariate exponential distribution. We assume only two nodes with exponential marginal distributions and then apply Crane and Hoek (2008) to find the closed form for conditional expectation formula for Gaussian copula. We derive the causal effect for the bivariate Gaussian copula. If we have a bivariate Gaussian copula, with dependence parameter ρ, we have

$$E(Y|X=x) = \int_{\mathbb{R}} y \frac{\partial}{\partial y} \Phi\left(\frac{\Phi^{-1}(F(y)) - \rho\Phi^{-1}(G(x))}{\sqrt{1-\rho^2}}\right) dy. \tag{12}$$

If both marginal distributions F and G were $N(0,1)$, the copula would revert back to the bivariate normal distribution. The Gaussian copula, however, gives us more flexibility, as it can accommodate any type of univariate distributions, F and G. In (12), we choose two marginal distributions that are exponential with parameter

		$\alpha = 0.01$		$\alpha = 0.1$		DAG known	
	n	**IDA**	**NCE**	**IDA**	**NCE**	**IDA**	**NCE**
p= 10							
	100	0.101	0.176	0.144	0.154	0.118	0.155
	1000	0.033	0.085	0.029	0.028	0.031	0.030
p= 50							
	100	3.732	2.515	2.261	1.759	2.004	1.677
	1000	1.175	1.100	0.964	0.378	0.724	0.281

Table 1: Comparison of mean absolute deviations (smaller is better) for causal effect estimates between our NCE and IDA (Maathuis et al., 2009) methods for small graphs ($p = 10$) and large graphs ($p = 50$) when the data is Gaussian.

$\lambda_x, \lambda_y > 0$. Thus, Equation (12) reduces to

$$\begin{aligned} E(Y|X=x) = & \frac{1}{\sqrt{1-\rho^2}} \\ & \int_{\mathbb{R}} y\phi\left(\frac{\Phi^{-1}(1-\exp(-\lambda_y y)) - \rho\Phi^{-1}(1-\exp(-\lambda_x x))}{\sqrt{1-\rho^2}}\right) \\ & \times \frac{\lambda_y \exp(-\lambda_y y)}{\phi(\Phi^{-1}(1-\exp(-\lambda_y y)))} dy, \end{aligned} \tag{13}$$

where $\phi(x) = \Phi'(x)$ is the standard normal density. Therefore, for a bivariate non-paranormal with exponential marginals, we obtain the following causal effect,

$$\begin{aligned} \mathrm{CE}(Y|X=x) = & -\frac{\rho}{1-\rho^2} \int_{\mathbb{R}} y\phi'(t) \frac{\lambda_x \exp(-\lambda_x x)}{\phi(\Phi^{-1}(1-\exp(-\lambda_x x)))} \\ & \times \frac{\lambda_y \exp(-\lambda_y y)}{\phi(\Phi^{-1}(1-\exp(-\lambda_y y)))} dy \end{aligned} \tag{14}$$

where $t = \frac{\Phi^{-1}(1-\exp(-\lambda_y y)) - \rho\Phi^{-1}(1-\exp(-\lambda_x x))}{\sqrt{1-\rho^2}}$.

In the simulation study we assume that node X affects node Y, in the following fashion,

$$\begin{aligned} X &= F^{-1}(\Phi(Z_1)) \\ Y &= F^{-1}\left(\Phi\left(\frac{Z_1 + Z_2}{\sqrt{2}}\right)\right), \end{aligned}$$

where F is the CDF of an Exponential(1) distribution and $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} N(0,1)$. This falls under the usual nonparanormal scenario. The explicit expression for the causal effect in Theorem 1 is very involved, but we derived in (14) a simplified expression. We evaluated this expression numerically to obtain the true causal effect, expressed as the solid black line in Figure 4. Then we simulated $n = 1,000$ observations from the above model for inferring the causal effect.

We assume that the underlying causal graph, $X \longrightarrow Y$, is known and used the NCE method to infer the non-linear causal effect. The blue line Figure 5.2 shows the functional causal effect estimate from NCE method. It matches very well the true causal effect. Clearly, had IDA been applied in this scenario, it would have come up with a nonsensical constant causal effect.

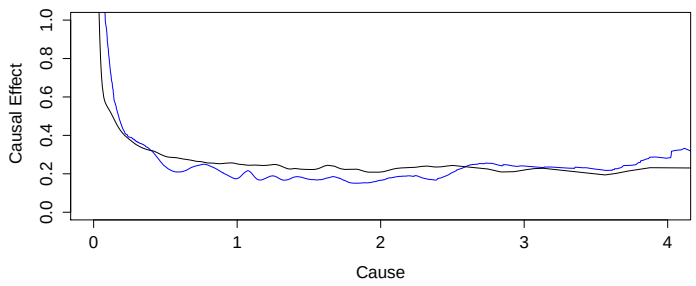


Figure 4: Exponential nonparanormal simulation: black line shows the true causal effect and the blue line represents the causal effect estimated by our NCE method.

5.3 Cauchy data

Although not the primary aim of our NCE method, it also consists of the RPC causal graph reconstruction method and can therefore be compared to the NPN method by Teramoto et al. (2014). We evaluated the performance of our NCE and the NPN methods for two different graph sizes: with 20 vertices and 200 vertices. We repeated this experiment 50 times for $n \in \{10, 100\}$.

We consider distributional systems whereby the underlying marginal distributions are mixtures of normal and Cauchy distributions. The mixing rate indicates the percentage of samples whose error distribution was drawn from the standard Cauchy distribution. We chose mixing rate 0.1, 0.5 and 1. The higher the mixing rate, the less accurate is the Gaussianity assumption. We averaged the values of true positive rate (TPR), false positive rate (FPR), true discovery rate (TDR) in the reconstruction of the causal graph. To compare NCE method with NPN method,

			TPR		FPR		TDR	
	n	**Mixing rate**	**NPN**	**NCE**	**NPN**	**NCE**	**NPN**	**NCE**
p= 20								
	10	0.1	0.412	0.432	0.00038	0.00027	0.887	0.899
		0.5	0.431	0.453	0.00025	0.00011	0.904	0.917
		1	0.514	0.524	0.00027	0.00016	0.911	0.929
	100	0.1	0.593	0.612	0.00033	0.00011	0.944	0.960
		0.5	0.654	0.671	0.00037	0.00018	0.958	0.971
		1	0.668	0.692	0.00041	0.00023	0.967	0.982
p= 200								
	10	0.1	0.311	0.323	0.00021	0.00005	0.891	0.911
		0.5	0.328	0.341	0.00032	0.0001	0.894	0.928
		1	0.337	0.348	0.00038	0.00014	0.902	0.934
	100	0.1	0.538	0.544	0.00017	0.00001	0.951	0.963
		0.5	0.562	0.581	0.00023	0.00006	0.963	0.977
		1	0.588	0.601	0.00027	0.0001	0.971	0.988

Table 2: Mean true positive, false positive and true discovery rates for the comparison of our NCE and NPN (Teramoto et al., 2014) methods for small graph ($p = 20$) and large graph ($p = 200$) when the data is Cauchy.

we show the representative results of setting for $\alpha = 10^{-4}$ in Table 2. It is clear that the NCE method, with its underlying RPC causal reconstruction method, always outperforms the NPN method.

6 TiMet: circadian regulation in *Arabidopsis Thaliana*

In this section, we illustrate our proposed approach by applying it to a time course gene expression dataset related to the study of circadian regulation in plants. The data used in our study come from the EU project TiMet (FP7-245143, 2014), whose objective is the elucidation of the interaction between circadian regulation and metabolism in plants.

The data consist of transcription profiles for the core clock genes from the leaves of various genetic variants of *Arabidopsis Thaliana*. The transcription profiles of the core clock genes (Aderhold, Husmeier, and Grzegorczyk, 2014, Pokhilko,

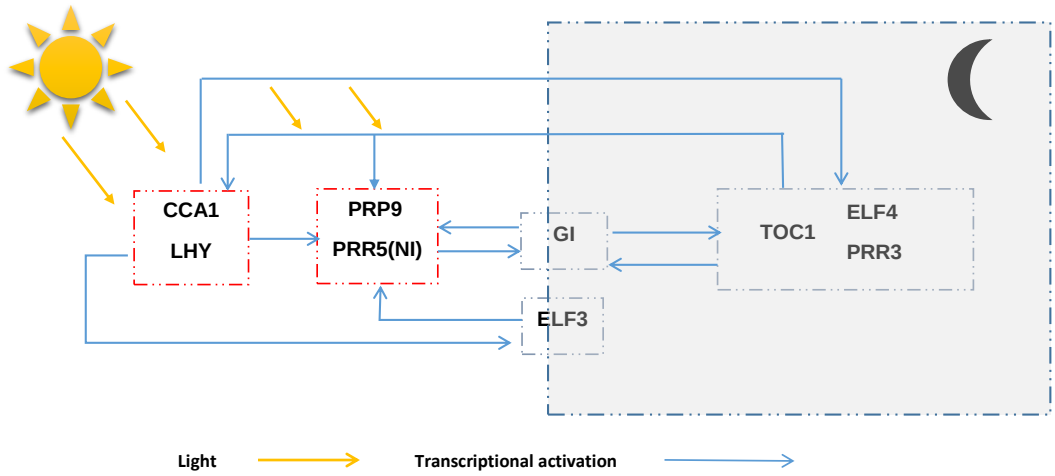


Figure 5: The inferred causal network among the circadian clock genes for *Arabidopsis thaliana*. The representation is inspired by Figure 1 in Jia and Huan (2009), showing significant overlap in topology.

Hodge, Stratford, Knox, Edwards, Thomson, Mizuno, and Millar, 2010, Guerriero, Pokhilko, Fernández, Halliday, Millar, and Hillston, 2012) were recorded: LHY, CCA1, PRR3, NI (PRR5), PRR9, TOC1, ELF3, ELF4 and GI. The plants were grown in the following 3 light conditions: a diurnal cycle with 12 hour light and 12 hour darkness (12L/12D), an extended night with full darkness for 24 hours, and an extended light with constant light for 24 hours. An exception is the ELF3 mutant, which was grown only in 12L/12D condition. Samples were taken every 2 hours to measure mRNA concentrations. We consider the same group of nine genes, which from previous studies are known to be involved in circadian regulation (Grzegorczyk and Husmeier, 2011a,b, Grzegorczyk, Husmeier, Edwards, Ghazal, and Millar, 2008, Jia and Huan, 2009). They consist of two groups of genes: "Morning genes", which are LHY, CCA1, PRR9, and PRR5, whose expression peaks in the morning, and "Evening genes", including TOC1, ELF4, ELF3, GI, and PRR3, whose expression peaks in the evening. The expressions for all the genes are strictly positive and highly right-skewed.

In traditional analysis of microarray studies, data are typically log-transformed. Especially when using the data for prediction, such transformations are sensible as they typically stabilize variances and make down-stream analyses more robust. In our case, however, our aim is to describe the system. We are *not* interested in the

causal effect of the log-transformed variables, but we are interested in the causal effects of the original variables. For this reason, we consider the raw data directly, since this is the scale on which we would like to evaluate the system.

For inferring the underlying causal CPDAG, we considered the RPC-algorithm in the version that uses the Kendall's tau — results using Spearman's rho were almost the same. The CPDAG contains three Markov equivalence DAGs. One of these three causal networks among the genes is displayed in Figure 5. For all three causal DAGs, we infer the causal effects between the genes and these are shown as lines in Figure 6. A striking feature is that most of the causal effects shrink towards zero for large values of the cause, possibly indicating saturation. Moreover, this effect seems stronger for the morning genes on the evening genes than vice versa.

The morning gene CCA1 was found to repress the evening genes EFL3 and NI. Among the evening genes, EFL4 and TOC1 have the strongest effect on both other evening and morning genes. The evening gene ELF positively affects CCA1. It also has a negative effect on LHY. Moreover, the evening genes ELF3, GI and TOC1 are involved in the activation of the morning gene PRR. The morning gene LHY has an almost constant effect on the evening genes ELF4, TOC1 and EFL4. In particular ELF4 interacts positively with NI and CCA1 and negatively with LHY. Many of these results are consistent with the findings in Grzegorczyk and Husmeier (2011a,b), Aderhold et al. (2014) and references therein.

Furthermore, we compare our network with the biological network referred to in Jia and Huan (2009), which is based on the work of Mas (2008) and Salome and McClung (2004). The most striking difference is that we have found evidence that the GI and ELF3 genes interact directly with the Pseudo-Response Regulators (PRR9 and PRR5 module). Chow, Helfer, Nusinow, and Kay (2012) suggest that this may be explained by the fact that another protein, LUX, which is not considered in this study, creates a complex with ELF3 that is required to regulate PRR9. This is an interesting methodological issue: studies based on lab-based pairwise interaction studies or studies looking for structural evidence of binding between proteins will not find links between proteins that are not directly interacting, whereas studies such as ours that perform network based analysis on a limited number of proteins will typically infer links between proteins that in reality require an intermediary not considered in the study.

7 Conclusion

In this paper, we have derived an explicit formula for causal effects in a flexible class of distributions, the so-called nonparanormal. These distributions are

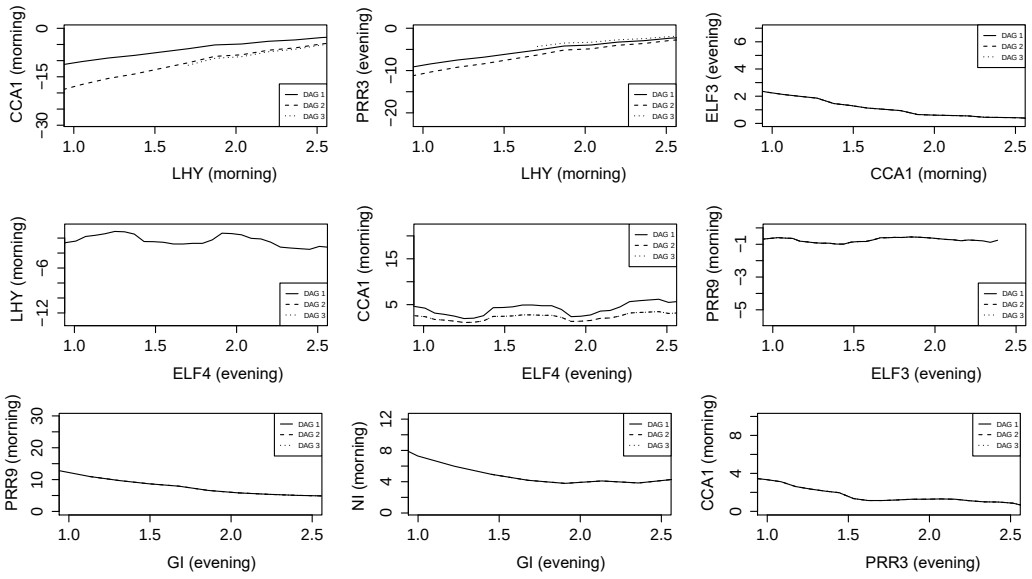


Figure 6: Causal effects for the circadian gene interaction network in *Arabidopsis thaliana*. Whereas ELF3 and ELF4 have almost constant causal effects, most of the others have a distinctive shrinkage in their causal effects for larger values of the cause, indicating saturation.

especially useful for real-life observational studies, where normality assumptions are often not warranted. We presented a simple method, NCE, to estimate these causal effects nonparametrically, based on a first order approximation of the general causal effect formula. It is able to capture a large range of non-linear causal effects and is shown to be consistent under certain conditions. In a simulation study, we have shown that the estimation method works well, particularly away from the tails of the data. We have also applied the method to an *Arabidopsis Thaliana* circadian clock network. The estimated causal effects reveal a tendency for some of the causal effects to shrink to zero for large values of the cause, which means that gene regulation shows effect saturation for high levels of the regulator. This is in correspondence with simple Michaelis-Menten kinetic models, often used to model gene regulation.

8 Appendix: Proofs

Proof of Theorem 1

Proof. We follow three steps for proving this theorem. First, we find a closed form expression for $E\left[Y|X_i = x_i; X_{\text{pa}(i)} = x_{\text{pa}(i)}\right]$. After that we connect this to the do-operator as is done in (2). Finally, taking the derivative in the way that the total causal effect is defined in (1) will complete the proof. From the differentiability of f_i it follows that the marginal distributions F_i are one-to-one, where $f_i^{-1}(x_i) = z_i$ and $Z_i = f_i^{-1}(X_i) = \Phi^{-1} \circ F_i(X_i)$ and $Z = f_y^{-1}(Y) = \Phi^{-1} \circ F_y(Y)$. Using the Taylor expansion,

$$\begin{aligned}
E\left[Y|X_i = x_i; X_{\text{pa}(i)} = x_{\text{pa}(i)}\right] &= E(F_y^{-1}(\Phi(Z))|X_i = x_i; X_{\text{pa}(i)} = x_{\text{pa}(i)}) \\
&= E(F_y^{-1}(\Phi(Z))|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)}) \\
&= E(f_y(Z)|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)}) \\
&= E(\sum_{k=1}^{\infty} f_y^{(k)}(z_0)\frac{(Z - z_0)^k}{k!}|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)}) \\
&= \sum_{k=1}^{\infty} f_y^{(k)}(z_0)\frac{1}{k!}E(Z^{*k}|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)}), \quad (15)
\end{aligned}$$

where $Z^* = Z - z_0$ for any $z_0 \in \mathbb{R}$. From the conditional normal distribution, we know that

$$Z^*|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)} \sim N(-z_0 + (\beta_i, \beta_{\text{pa}(i)})(z_i, z_{\text{pa}(i)})^T, (1 - \rho^2)).$$

where $(\beta_i, \beta_{\text{pa}(i)}) = \Sigma_{p,(i,\text{pa}(i))}\Sigma^{-1}_{(i,\text{pa}(i)),(i,\text{pa}(i))}$ and $\rho = \Sigma_{p,(i,\text{pa}(i))}\Sigma^{-1}_{(i,\text{pa}(i)),(i,\text{pa}(i))}\Sigma_{(i,\text{pa}(i)),p}$.

Following Lehmann and Casella (1998) page 132, we get for $k \in \mathbb{N}$

$$\begin{aligned}
E(Z^{*k}|Z_i = z_i; Z_{\text{pa}(i)} = z_{\text{pa}(i)}) = \sum_{r=0}^{\lfloor \frac{k}{2} \rfloor} \binom{k}{2r} (-z_0 + \beta_i z_i + \beta_{\text{pa}(i)}^T z_{\text{pa}(i)})^{k-2r} \\
\times (2r-1)\ldots 3 \times 1 \times [(1-\rho^2)]^r. \quad (16)
\end{aligned}$$

Plugging (16) into (15), we have

$$\begin{aligned}
E(Y|X_i = x_i; X_{\text{pa}(i)} = x_{\text{pa}(i)}) = \sum_{k=1}^{\infty}\sum_{r=0}^{\lfloor \frac{k}{2} \rfloor} f_y^{(k)}(z_0)\frac{1}{k!}\binom{k}{2r} \\
\times (-z_0 + \beta_i z_i + \beta_{\text{pa}(i)}^T z_{\text{pa}(i)})^{k-2r} \\
\times (2r-1)\ldots \times 3 \times 1 \times [(1-\rho^2)]^r. \quad (17)
\end{aligned}$$

Now we use (17) for finding the intervention effect for nonparanormal variable. That is,

$$
\begin{aligned}
E(Y|\mathrm{do}(X_i = x_i)) &= \int E(Y|X_i = x_i; X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)}) P(x_{\mathrm{pa}(i)})\, d(x_{\mathrm{pa}(i)}) \\
&= \sum_{k=1}^{\infty} \sum_{r=0}^{\lfloor \frac{k}{2} \rfloor} f_y^{(k)}(z_0) \frac{1}{k!} \binom{k}{2r} (2r-1)\ldots 3.1 \times [(1-\rho^2)]^r \\
&\times \sum_{s=0}^{k-2r} \binom{k-2r}{s} (-z_0 + \beta_i z_i)^s \\
&\times \int (\beta_{\mathrm{pa}(i)}^T z_{\mathrm{pa}(i)})^{k-2r-s} P(z_{\mathrm{pa}(i)})\, d(z_{\mathrm{pa}(i)}) \\
&= \sum_{k=1}^{\infty} \sum_{r=0}^{\lfloor \frac{k}{2} \rfloor} f_y^{(k)}(z_0) \frac{1}{k!} \binom{k}{2r} \times (2r-1)\ldots 3.1 \times [(1-\rho^2)]^r \\
&\times \sum_{s=0}^{k-2r} \binom{k-2r}{s} (-z_0 + \beta_i z_i)^s E[(\beta_{x_{\mathrm{pa}(i)}}^T Z_{\mathrm{pa}(i)})^{k-2r-s}]. \qquad (18)
\end{aligned}
$$

We get the following expression for the total causal effect,

$$
\frac{\partial}{\partial x_i} E[Y|\mathrm{do}(X_i = x_i)] = \frac{\partial}{\partial z_i} E[Y|\mathrm{do}(X_i = x_i)] \frac{\partial z_i}{\partial x_i}, \qquad (19)
$$

where $\frac{\partial z_i}{\partial x_i} = (f_i^{-1})'(x_i)$. Therefore, with plugging (18) into (19), the proof is completed. □

Proof of Proposition 3

Proof. Define

$$
\hat{Z}_n = \begin{pmatrix}
\hat{z}_{1,i} & \hat{z}_{1,\mathrm{pa}(i)_1} & \cdots & \hat{z}_{1,\mathrm{pa}(i)_k} \\
\hat{z}_{2,i} & \hat{z}_{2,\mathrm{pa}(i)_1} & \cdots & \hat{z}_{2,\mathrm{pa}(i)_k} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{z}_{N,i} & \hat{z}_{N,\mathrm{pa}(i)_1} & \cdots & \hat{z}_{N,\mathrm{pa}(i)_k}
\end{pmatrix},
$$

such that $\hat{z}_{j,l} = \Phi^{-1}(\hat{F}_{l,\mathrm{n}}(x_{jl}))$ where x_{jl} is the non-ordered jth sample of variable l and $\mathrm{pa}(i)$ is the index set of k parents of i.Let

$$
\hat{\Upsilon}_n^T = \left(\Phi^{-1}(\hat{F}_{y,\mathrm{n}}(y_1)), \Phi^{-1}(\hat{F}_{y,\mathrm{n}}(y_2)), \cdots, \Phi^{-1}(\hat{F}_{y,\mathrm{n}}(y_N))\right).
$$

The coefficient $\hat{\beta}_i^n$ is defined as the first element of the vector,

$$
\hat{\beta}^n = (\hat{Z}_n^t \hat{Z}_n)^{-1} \hat{Z}_n^t \hat{\Upsilon}_n.
$$

We can also define the oracle estimator $\hat{B}_i^n$ as the first element of

$$\hat{B}^n = (Z_n^t Z_n)^{-1} Z^t \Upsilon_n,$$

where Z_n and Υ_n are obtained by replacing the marginal $\hat{F}$s by the true Fs. Consider an arbitrary $\varepsilon, \delta > 0$,

$$\begin{aligned} P(|\hat{\beta}_i^n - \beta_i| > \varepsilon) &= P(|\hat{\beta}_i^n - \hat{B}_i^n + \hat{B}_i^n - \beta_i| > \varepsilon) \\ &\leq P((|\hat{\beta}_i^n - \hat{B}_i^n| + |\hat{B}_i^n - \beta_i|) > \varepsilon) \\ &\leq P((|\hat{\beta}_i^n - \hat{B}_i^n| > \varepsilon/2) + P(|\hat{B}_i^n - \beta_i|) > \varepsilon/2). \end{aligned} \tag{20}$$

We first consider the first right hand side term of (20). Let's define $\hat{A}_n = \frac{\hat{Z}_n^t \hat{Z}_n}{n}$, $A_n = \frac{Z_n^t Z_n}{n}$ and $\hat{b}_n = \frac{\hat{Z}_n^t \hat{\Upsilon}_n}{n}$ and $b_n = \frac{Z_n^t \Upsilon_n}{n}$. Then,

$$\begin{aligned} P(|\hat{\beta}_i^n - \hat{B}_i^n| > \frac{\varepsilon}{2}) &\leq P(\| \hat{A}_n^{-1} \hat{b}_n - A_n^{-1} b_n \|^2 > \frac{\varepsilon}{2}) \\ &\leq P(\| \hat{A}_n^{-1} (\hat{b}_n - b_n) \|^2 \\ &\quad + \| (\hat{A}_n^{-1} - A_n^{-1}) b_n \|^2 > \frac{\varepsilon}{2}) \\ &\leq P(\| \hat{A}_n^{-1} (\hat{b}_n - b_n) \|^2 > \frac{\varepsilon}{4}) \\ &\quad + P(\| (\hat{A}_n^{-1} - A_n^{-1}) b_n \|^2 > \frac{\varepsilon}{4}). \end{aligned} \tag{21}$$

By the consistency of $\hat{z}$, we have that both $\hat{b}_n$ and b_n converge in probability to some $b = \Sigma_{(i,\mathrm{pa}(i)),p}$ and both $\hat{A}_n^{-1}$ and A_n^{-1} converge in probability to some $A^{-1} = \Sigma^{-1}_{(i,\mathrm{pa}(i)),(i,\mathrm{pa}(i))}$, where Σ is defined in the body of Theorem 1. Therefore, there is a n^*, such that for all $n \geq n^*$, both terms on the right hand side of (21) are less than $\delta/4$. So for all $n \geq n^*$,

$$P(|\hat{\beta}_i^n - \hat{B}_i^n| > \frac{\varepsilon}{2}) < \frac{\delta}{2}.$$

For the second term of the right hand side of (20), it is sufficient to use the fact that in the latent normal space a regression estimate is consistent and therefore, there exist a $n^\perp$, such that any $n > n^\perp$,

$$P(|\hat{B}_i^n - \beta_i| > \varepsilon/2) < \delta/2.$$

Putting both results together, we now have that for any $n \geq \max\{n^*, n^\perp\}$,

$$P(|\hat{\beta}_i^n - \beta_i| > \varepsilon) < \delta.$$

Thus we get the desired result. □

Proof of Proposition 4

Proof. For two sequences of random variables Z_n and W_n and two random variables Z, W, such that Z_n converges in probability to Z and W_n converges in probability to W, then it is a standard result that $Z_n W_n$ converges in probability to ZW (Lehmann, 2004). As all the components of $\widehat{\text{NCE}}_0(x)$ have been shown to be consistent, then the estimator is consistent. □

References

Aderhold, A., D. Husmeier, and M. Grzegorczyk (2014): "Statistical inference of regulatory networks for circadian regulation," *Statistical applications in genetics and molecular biology*, 13, 227–273.

Anderson, T. (2003): *An introduction to multivariate statistical analysis*, Wiley series in probability and statistics,New YorkChichester: Wiley.

Chickering, D. M. (2002): "Learning equivalence classes of bayesian-network structures," *The Journal of Machine Learning Research*, 2, 445–498.

Chickering, D. M. (2003): "Optimal structure identification with greedy search," *The Journal of Machine Learning Research*, 3, 507–554.

Chow, B. Y., A. Helfer, D. A. Nusinow, and S. A. Kay (2012): "Elf3 recruitment to the prr9 promoter requires other evening complex members in the arabidopsis circadian clock," *Plant signaling & behavior*, 7, 170–173.

Crane, G. J. and J. v. d. Hoek (2008): "Conditional expectation formulae for copulas," *Australian & New Zealand Journal of Statistics*, 50, 53–67.

Grzegorczyk, M. and D. Husmeier (2011a): "Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes," *Bioinformatics*, 27, 693–699.

Grzegorczyk, M. and D. Husmeier (2011b): "Non-homogeneous dynamic bayesian networks for continuous data," *Machine Learning*, 83, 355–419.

Grzegorczyk, M., D. Husmeier, K. D. Edwards, P. Ghazal, and A. J. Millar (2008): "Modelling non-stationary gene regulatory processes with a non-homogeneous bayesian network and the allocation sampler," *Bioinformatics*, 24, 2071–2078.

Guerriero, M. L., A. Pokhilko, A. P. Fernández, K. J. Halliday, A. J. Millar, and J. Hillston (2012): "Stochastic properties of the plant circadian clock," *Journal of The Royal Society Interface*, 744–756.

Gugushvili, S. and C. A. J. Klaassen (2012): "$\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing," *Bernoulli*, 18, 1061–1098.

Harris, N. and M. Drton (2013): "Pc algorithm for nonparanormal graphical models," *The Journal of Machine Learning Research*, 14, 3365–3383.

Heckerman, D. and D. Geiger (1995): "Learning bayesian networks: a unification for discrete and gaussian domains," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 274–284.

Jia, Y. and J. Huan (2009): "The analysis of arabidopsis thaliana circadian network based on non-stationary dbns approach with flexible time lag choosing mechanism," in *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE International Conference on*, IEEE, 178–181.

Kalisch, M. and P. Bühlmann (2007): "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *The Journal of Machine Learning Research*, 8, 613–636.

Lehmann, E. L. (2004): *Elements of large-sample theory*, Springer Science & Business Media.

Lehmann, E. L. and G. Casella (1998): *Theory of point estimation*, volume 31, Springer Science & Business Media.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012): "High-dimensional semiparametric gaussian copula graphical models," *The Annals of Statistics*, 40, 2293–2326.

Maathuis, M. H., M. Kalisch, and P. Bühlmann (2009): "Estimating high-dimensional intervention effects from observational data," *The Annals of Statistics*, 37, 3133–3164.

Mas, P. (2008): "Circadian clock function in arabidopsis thaliana: time beyond transcription," *Trends in cell biology*, 18, 273–281.

Mooij, J. M., D. Janzing, T. Heskes, and B. Schölkopf (2011): "On causal discovery with cyclic additive noise models," in *Advances in neural information processing systems*, 639–647.

Nandy, P., M. H. Maathuis, and T. S. Richardson (2017): "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings," *The Annals of Statistics*, 45, 647–674.

Pearl, J. (1995): "Causal diagrams for empirical research," *Biometrika*, 82, 669–688.

Pearl, J. (2009): *Causality*, Cambridge university press.

Pokhilko, A., S. K. Hodge, K. Stratford, K. Knox, K. D. Edwards, A. W. Thomson, T. Mizuno, and A. J. Millar (2010): "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," *Molecular systems biology*, 6, 416.

Priestley, M. and M. Chao (1972): "Non-parametric function fitting," *Journal of the Royal Statistical Society. Series B (Methodological)*, 385–392.

Richardson, T. (1996): "A discovery algorithm for directed cyclic graphs," *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, 454–461.

Salome, P. A. and C. R. McClung (2004): "The arabidopsis thaliana clock," *Journal of Biological Rhythms*, 19, 425–435.

Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen, and R. G. Cowell (1993): "Bayesian analysis in expert systems," *Statistical science*, 219–247.

Spirtes, P., C. N. Glymour, and R. Scheines (2000): *Causation, prediction, and search*, volume 81, MIT press.

Spirtes, P., C. Meek, and T. Richardson (1995): "Causal inference in the presence of latent variables and selection bias," *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 499–506.

Teramoto, R., C. Saito, and S.-i. Funahashi (2014): "Estimating causal effects with a non-paranormal method for the design of efficient intervention experiments," *BMC bioinformatics*, 15, 228.

Verma, T. and J. Pearl (1990): "Equivalence and synthesis of causal models [technical report r-150]," *Department of Computer Science, University of California, Los Angeles*.